DOCUMENT RESUME

ED 097 375 95 TM 004 017

AUTHOR Frick, Ted; Semmel, Melvyn I.
TITLE Observational Records: Observer Agreement and Reliabilities.
INSTITUTION Indiana Univ., Bloomington. Center for Innovation in Teaching the Handicapped.
SPONS AGENCY Bureau of Education for the Handicapped (DHEW/OE), Washington, D.C. Div. of Research.
PUB DATE [Apr 74]
GRANT OEG-9-242178-4149-032
NOTE 54p.; Paper presented at the Annual Meeting of the American Educational Research Association (59th, Chicago, Illinois, April 1974)

EDRS PRICE MF-$0.75 HC-$3.15 PLUS POSTAGE
DESCRIPTORS Classroom Observation Techniques; *Observation; *Reliability; *Statistical Analysis; *Student Behavior; *Teacher Behavior

ABSTRACT

Observer disagreement is important insofar as it limits the reliabilities of observational records. This discussion evolves around methods and conditions under which observer agreement can be measured as to minimize such an occurrence. Observers should be trained to nearly perfect agreement with a criterion or expert coder on unambiguous examples of behavioral categories before actual data collection. Disagreement on ambiguities may help reflect a more accurate representation of the real world. In addition to criterion-related agreement, it is suggested that intraobserver agreement be obtained by showing a video tape twice to all observers in which conditions parallel those encountered in the field. While criterion-related and intraobserver agreement measures have been recommended for both before and during a study, they should not be used as evidence of observer agreement in the actual classroom, but rather to assist an investigator in documenting adequacy of observational skills. After a study is finished, reliabilities of observational data and coefficients of stability and observer agreement should be calculated by using intraclass correlation coefficients. (Author/RC)

# OBSERVATIONAL RECORDS:

## Observer Agreement and Reliabilities[1]

Ted Frick and Melvyn I. Semmel



Center for Innovation in Teaching the Handicapped

Indiana University

Presented at the 1974 meeting of the
American Educational Research Association
Chicago, Illinois, April 16, 1974

[1]This research was partially supported by the Intramural Research Program, Division of Research, Bureau for the Handicapped, U.S.O.E. through grant #OEG-9-242178-4149-032 to the Center for Innovation in Teaching the Handicapped. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

While there is a plethora of observation systems currently employed in research in teacher behavior (Simon & Boyer, 1970), there is a corresponding paucity of evidence relating specific teaching skills to changes in pupil classroom behavior (Smith, 1971). One possible reason for many insignificant findings may be that investigators have inadequately controlled for a number of sources of error associated with observational data (McGaw, Wardrop, & Bunda, 1972).

Prevalent confusion concerning reliabilities of observational records can be traced to failure of separating two statistically related but conceptually different measures: observer agreement and reliabilities of observational records (Medley & Mitzel, 1958; 1963; McGaw, et al., 1972). It is generally agreed that the reliability of a test is a necessary but not a sufficient condition for determining concurrent or predictive validity. Analogously, observer agreement is a primary issue, although not the most important one, to be faced in interpretation of results of an observational study. More important are the reliabilities of the observational data--i.e., the extent to which observational records discriminate teachers, pupils, and situations within classroom environments. Observer disagreement is important insofar as it acts as a limiting factor on reliabilities of observational data.

Reliabilities of teacher/pupil behavior and other classroom process variables are especially important in studies attempting to relate these variables to outcome measures such as pupil growth (Soar, 1972). Unreliability of either or both measures will tend to obscure any significant relationships that may exist. Moreover, reliabilities of classroom process variables are ultimately essential for the purpose of generalization about relationships among teacher and pupil behaviors. Given that certain

pupil outcomes are predictably associated with specific teaching strategies, then such information can be communicated to teachers for use in daily decision-making.

In dealing with traditional measuring instruments, reliability has been classically defined as the consistency with which the instrument measures something. Typically this has been done by using parallel forms of a test, test-retest methods, or methods of internal consistency. A number of assumptions, however, are made about this parallel measures concept. That is, two or more tests are assumed to be equivalent in content, means, variance, and intercorrelations of items (Cronbach, Rajaratnam, & Gleser, 1963).

While these assumptions are seldom fully met in practice when using traditional tests, they become impractical when applied to observational studies where human raters (who take the place of tests) are rarely identical or equivalent in their observational skills.

In order to avoid such assumptions a number of statisticians have proposed the use of intraclass correlation coefficients as a means of determining reliabilities (Haggard, 1958; Cronbach, et al., 1963; Gleser, Cronbach, & Rajaratnam, 1965; Medley & Mitzel, 1958, 1963; McGaw, et al., 1972). There has been general agreement about the formula for a reliability in terms of population parameters.

That is:

$$\rho_{xx} = \frac{\sigma_t^2}{\sigma_x^2} = \frac{\sigma_t^2}{\sigma_t^2 + \sigma_\varepsilon^2}$$

where:

$\sigma_t^2$ is defined as the variance of the true scores (e.g., of teacher behavior) around the mean of all such true scores in the population of teachers represented by the sample of teachers actually observed.

And:

> $\sigma^2_X$ is not as easily defined. The definition of $\sigma^2_X$ will vary according to the procedure used to collect the data. Basically, $\sigma^2_X$ represents the variance of the obtained measures on all the teachers in the population about their own mean. (Medley & Mitzel, 1963).

While specific components of variance included in $\sigma^2_X$ are defined by the design of a study and an investigator's particular interests, the obtained ~~variance~~ can be considered to consist of a true variance component, $\sigma^2_t$, and a generic error variance component, $\sigma^2_\varepsilon$ (McGaw, et al., 1972). Likewise, some components of generic error variance are determined by the nature of the reliability coefficient of interest. However, observer disagreement, or error variance attributable to observers, is almost always considered to be a part of the generic error variance when computing reliabilities of observational records. Thus, minimal observer disagreement is a necessary but insufficient condition for high reliability coefficients, since there are other components of the generic error variance which are theoretically independent from observer error variance (e.g., intra-subject variance from occasion to occasion).

Other sources of error can contribute to unreliability of observational records as well as observer disagreement. Instability of teacher/pupil behavior from occasion to occasion is typically the greatest source of error (Medley & Mitzel, 1963). Poorly designed observation systems and studies can also contribute to unreliability. Finally, inappropriate data analysis procedures can obfuscate actual differences among classrooms. Although human behavior is frequently unstable across separate occasions, it is possible to minimize observer errors and investigator errors.

## Observer Errors

Observer agreement is typically calculated by comparing observational records of two or more observers with each other or with an expert when simultaneously coding the same classroom events. Observer agreement is, however, not synonymous with reliabilities of observational records as is often mistakenly assumed. For example, observers can be trained to a high level of agreement, yet they can collect very unreliable data if the behaviors of the observed teachers/pupils differ little, or if behaviors are truly unstable from occasion to occasion. That is, if variance between subjects (true variance) is small relative to variance within subjects (error variance), the measurement will be unreliable regardless of the extent of observer agreement.

Nonetheless, observer disagreement cannot be totally ignored. If it cannot be documented that observers adequately agreed on the same behavioral events upon completion of training, and if the data they collected proved to be unreliable, a critical paradox is faced: Is the source of unreliability of observational records due primarily to lack of observer agreement or consistency in the field? Or, is the source of unreliability largely due to a lack of discriminable differences among teachers/pupils--either because behaviors are too unstable or because there are no existing differences among subjects?

If adequate observer agreement cannot be demonstrated, observational data may be confounded with errors attributable to observer misunderstandings of category definitions, inconsistencies, and biases which are realistically inseparable from actual inconsistencies (error variance) within teachers/pupils and true variation among teachers/pupils in classroom observation research designs.

Once a study is finished, the extent to which observer errors detract from reliabilities can be estimated by intraclass correlation coefficients, providing certain design requirements are met (Medley & Mitzel, 1963). It is too late, however, at this point in time to retrain observers if observer disagreement is found to be seriously limiting reliabilities. Hence, methods and conditions under which observer agreement can be measured so as to minimize such an occurrence are of prime concern.

Observer Agreement When?

Perhaps the first issue to be addressed is when measures of observer agreement can be made so as to minimize the possibility that it is observers who are the prime source of potentially unreliable observational data.

This suggests that a researcher should demonstrate that observers were adequately trained before actual data collection. Even if this prior agreement check was accomplished under conditions identical to those of actual data collection,[2] it is still no guarantee that observer skills will not deteriorate during data collection. It follows then that checks be made to insure that observers are maintaining their skills during the study as well. If certain observers are found to be deteriorating in their skills at an early stage in the study, remediation or deletion is possible without the loss of the entire study.

The frequency of maintenance tests largely depends on how much a researcher is willing to gamble against the possibility of obtaining results which are unreliable primarily because of observer disagreement. Of course, if results are reliable, observer agreement becomes a moot point unless there is a reason to suspect that observers are biased in some manner or that a "halo effect" may be contributing to reliability (see Medley & Mitzel, 1963).

[2]The authors do not intend to imply that this is the most desirable means of measuring observer agreement, as will be discussed later.

The number of maintenance trials also depends on other factors such as the complexity of the observation system, the design of the study, the contingencies under which observers code, the length of the study, the frequency of observations for each observer, scheduling problems, economic considerations, etc. It is wise to schedule a maintenance check some time early in the study in order to reduce the likelihood of collecting large amounts of data ridden with observer errors. At minimum, an additional test near the end of the study is recommended.

Agreement on What Kinds of Data?

Medley (personal communication, 1972) has noted that one logical consideration for determining observer agreement is that of how the data are to be analyzed. Although this may appear obvious, it is sometimes overlooked. For example, Flanders (1967) calculates observer agreement using a modification of Scott's $\pi$ index (1955) based on total frequencies across all categories, comparing two observers at a time. Yet in data analysis he utilizes two-stage behavior chains in his matrices. It is conceivable that observers could agree very well in overall category totals, yet profoundly disagree on identification of two-stage behavior patterns.

If data are to be analyzed by frequencies of individual categories emitted, it is suggested that observer agreement measures be determined on total frequencies for each category. If comparisons of frequencies of groups of categories (scales) are planned, it is suggested that observer agreement measures be based on total scale frequencies. Or, if analysis of three-stage patterns or chains of behavior is intended, computation of observer agreement on the basis of three-stage chains is recommended.

In short, observer agreement should be computed on the same unit(s) of behavior that will be used in later analyses.

Agreement with Whom?

As was stated earlier, observer agreement is typically calculated by comparing observational records of two or more observers with each other or with an "expert" when coding the same classroom events at the same time. Assuming that an analysis based on categorical frequencies was of interest, an interobserver agreement estimate (across all observers) could be obtained for each category by using intraclass correlation coefficients (Ebel, 1951; Haggard, 1958; Medley & Mitzel, 1958, 1963).

Suppose a coefficient of .85 or greater for each category is considered to be satisfactory evidence of adequate observer training. Aside from problems of interpretation which will be discussed later, other considerations are paramount. What if coefficients are much lower than .85 on several of the categories? Does this mean that these categories are not clearly defined? Or does it mean that some of the observers do not clearly understand them? If so, which ones are the bad coders? Are they the ones who deviate most widely from the rest of the group? Conversely, are these "deviates" the good coders, and the remainder of the group making a common mistake? Moreover, even on categories with coefficients greater than .85, such high intraobserver agreement does not necessarily reflect high agreement with the original definitions of these categories.

In attempting to reduce the above-mentioned problems of using an interobserver agreement measure, it seems advantageous to compare each observer's scores to those of a criterion or an expert coder with whom the researcher has confidence for strictly, consistently, and objectively following original category definitions. This type of observer agreement is referred to as criterion-related agreement. Such an agreement measure is more useful than an interobserver agreement measure when decisions about adequacy of individual observer skills are paramount.

Moreover, criterion-related observer agreement lends itself well to the design of instructional materials for observer training (Thiagarajan, 1973). One of the problems in drawing conclusions from observational studies relating teacher behavior to pupil growth is that it is difficult to generalize findings from independent studies in which information about the behavioral variables is inadequate (Soar, 1972). Criterion-referenced observer agreement measures within the context of an observation system instructional package could help reduce this problem.

Agreement Under What Conditions and How "Perfect?"

It has been implied that perfect observer agreement is desirable, but the conditions under which this applies have not been specified, nor have the conditions been explicated under which observer disagreement is desirable.

Medley and Norton (1971) have purported that studies performed in the field to determine observer agreement before actual observational data collection should be discontinued. Rather, investigators need only document that their observers were competent upon completion of training--competency being determined via nearly perfect observer agreement on unambiguous examples of behavioral categories shown on video tape. Conversely, they have argued that perfect observer agreement during actual data collection may not be particularly desirable. Since teachers and pupils in the real world do not always exhibit behaviors that neatly fall into predefined observational system categories, observer disagreement on ambiguities reveals a more representative picture of that real world.

Although their argument is initially disconcerting in that it may appear inconsistent, it does make sense from a practical standpoint. It is highly improbable that any observation system has such specifically defined

and perfectly mutually exclusive categories that every behavioral event that occurs can be clearly fitted into one of its categories. In all likelihood there will be some teacher/pupil behaviors which are ambiguous--i.e., they contain elements of two or more categories in the system. If observers are brainwashed to the point that they consistently code the same ambiguous behaviors into a certain category, results could be biased. Alternatively, if one observer codes an ambiguous behavior into one category and another observer codes the same ambiguous behavior into a different category, the overall results may indicate a more realistic description of that teacher's behavior. That is, in the latter case there will be some tallies in both categories, rather than in only one category as in the former case.

Medley and Norton (1971) have suggested that a videotape containing solely unambiguous examples of each category for measuring observer agreement should be constructed. This could be done by taping a variety of live classrooms. A panel of judges could then review these tapes and select a number of isolated, unambiguous examples to be dubbed onto another tape. This process, however, can be quite time-consuming and uneconomical. Moreover, the video and audio quality of such tapes tends to be rather poor. It is also sometimes difficult to find an adequate number of examples of certain behavioral categories which occur infrequently in actual classrooms.

An alternative approach is to videotape classroom simulations of isolated examples. This can be done in an environment where acceptable audio-visual tape quality can be maintained. Furthermore, examples can be randomly ordered, while producing only one original master tape with segment signs, directions, etc., concurrently recorded.

The issue that this type of observer agreement measure lacks validity could be raised. That is, observers are not being tested under the conditions

to which they are exposed during action observation. While this is true, it should be noted that the purpose of such a criterion-related agreement measure is to test an observer's knowledge of the items on the observation instrument. If observer agreement is measured under actual observation conditions, it is usually impossible to separate unwanted observer errors in coding unambiguous examples from expected and desirable observer disagreement on ambiguous ones. Therefore, it will be difficult to establish an acceptable level of agreement for making decisions about individual observer competencies.

As a compromise to this issue of validity, observers can also be tested under more realistic conditions in addition to that of coding isolated unambiguous examples. Because of the above-mentioned paradox, however, it is suggested that observer agreement measures be interpreted in a different manner. While nearly perfect observer agreement with a criterion is expected for clear-cut examples, a different method and standard is recommended for measures taken under actual coding conditions.

The proposed additional method is that of showing a video tape of realistic conditions twice to all observers. This tape should probably be about the same length as a normal observation period in a study, contain numerous examples of each category, and be fairly representative of the conditions under which actual observation takes place. Rather than focusing on agreement with a criterion or with other observers, the issue of individual observer consistency can be addressed. The extent to which each observer is consistent with himself can be measured by comparing results from the first viewing to those of the second. Biases in coding ambiguous events that exist for each coder are expected to remain fairly consistent from viewing to viewing, and thus moderately high intraobserver agreement or consistency is demanded.

Again, a videotape simulation of classroom situations is advantageous, since it is usually easier to control audio-visual quality and provide an adequate number of examples of each category than it is in taping live classrooms.

While this paper stresses methods of making decisions about adequacy of observer training, observer vigilance is an equally important problem. Although observers can be trained quite well with carefully designed instructional packages (Thiagarajan, 1973) or with a computer-assisted consensus coding schema (Semmel, 1972), there is no guarantee that those who perform well on a criterion test or demonstrate high intraobserver agreement scores will perform well as coders in the field. Strategies for coping with observer vigilance lie beyond the scope of this paper. However, lack of observer vigilance can be a significant source of error that limits reliabilities of observational records.

How Can Agreement Be Measured?

Since there is probably no one best way of calculating observer agreement, a number of alternatives will be discussed. It should be noted that measures of observer agreement will be considered only in relation to criterion-related and intraobserver agreement checks. In the former an individual observer's codes on unambiguous, isolated examples are compared to a criterion in order to make decisions about the observer's knowledge of the system. In the latter test, an individual observer's codes from one viewing of a videotape of realistic classroom conditions are compared to the same observer's codes from a second viewing of the same videotape. In both types of agreement situations only two observers are compared at a time.

Intraclass Correlation Coefficients. Previous mention was made concerning the use of intraclass correlation coefficients for determining

interobserver agreement (Ebel, 1951; Medley & Mitzel, 1958). While it was pointed out that measures across all observers are impractical for making decisions about individual observers, one other problem is encountered with this method.

An intraclass correlation coefficient is an estimate of the ratio of true variance of behaviors to that of obtained variance. Obtained variance includes the true variance plus variance due to measurement error and any other factors not taken into consideration in the design (which are included in the error variance). If error variance attributable only to observers is isolated, and if the remaining variance is considered as true variance, the extent to which observer disagreement detracts from reliability of observed behaviors can be estimated (Medley & Mitzel, 1958).

The major problem faced in measuring observer agreement in this manner (outside of regular data collection) is that the amount of true variance fluctuates depending on whether or not the situations selected to be coded are grossly different or highly similar. If the selected situations differ widely on an observation category, then more observer error variance can be tolerated than for situations across which there is little variance in the category.

For example, suppose that an investigator constructs a videotape test with widely varying situations, gives it to two observers, and finds an estimated true variance of 90 (among situations on category X) and an observer error variance of 30. Then $r_{xx} = \frac{90}{90+30} = .75$. On the other hand, suppose that a test is constructed with highly similar situations (i.e., little variance among situations on category X) and is given to the same two observers. If the true variance was estimated to be 6 and the observers were consistent with their disagreement in the former example (error variance of 30), then $r_{xx} = \frac{6}{6+30} = .17$.

All other things being equal, the discrepancy between the two coefficients in the above example is due to the fluctuation in true variance among situations. Unless several previous studies with the same population and the same observation system have been done, it is difficult to estimate legitimately amounts of true variances expected in the field for teachers, pupils, and situations. Thus, it is hard to decide how much observer variance can be tolerated.

Moreover, even if the amount of expected true variance were known, it would probably be challenging to construct a criterion tape containing only unambiguous examples that has a true variance roughly equivalent to the expected true variance in the field.

Although intraclass correlation coefficients should be used to determine the extent to which observer disagreement limits reliabilities after a study is completed, they are impractical for determining individual observer competencies before a study is begun.

Simple Percentage Agreement.[3] Simple percentage agreement can be calculated in one of two ways: Two sets of ratings can be compared on an item by item basis. That is, on a given item the observer either agrees (A) or disagrees (D) with the criterion. Observer agreement for a particular category i (wherever the expert has recorded an i) is defined:

$$P_{o_i} = \frac{\Sigma A_i}{\Sigma A_i + \Sigma D_i}$$

Where $\Sigma A_i$ = total number of agreements for the ith category

and $\Sigma D_i$ = total number of disagreements for the ith category

[3]This and the following observer agreement measures are illustrated by example in Appendix A.

Agreement can also be calculated by comparing the two observers' total frequencies for a given category across a number of situations. That is,

$$P'_{o_i} = \frac{f_{1i}}{f_{2i}} \text{ or } \frac{f_{2i}}{f_{1i}}, \text{ such that } 0 \leq P'_{o_i} \leq 1.00$$

where $f_{1i}$ = total frequency for the observer on category $i$

and $f_{2i}$ = total frequency for the expert on category $i$

Notice that the first coefficient tends to be more stringent than the second. When using the former, an observer must be correct on almost every item (event) in order to achieve nearly perfect agreement, while on the second he could disagree on some specific items, yet end with total category frequencies very similar to the expert's.

It can be argued that the second method is more appropriate if data are to be analyzed on the basis of total category frequencies. Alternatively, high agreement when using the first method would almost assure that the observer really knew the system. One limitation of the first method, however, is that it is not always possible to compare ratings on an item by item basis.

There are two other disadvantages to using either of these methods of simple percentage agreement. When low frequencies of some categories occur while other categories occur very often, interpretation of coefficients may be ambiguous. For example, if an observer gets 2 out of 3 correct on category X, and 19 out of 20 correct on category Y, coefficients of .67 and .95 are respectively obtained. Yet the observer deviates by only one tally in each case.

A solution to this problem is to structure the criterion test so that it contains an approximately equal number of unambiguous examples of each category. It is suggested that 10 or more examples of each category be given in order to reduce the likelihood that an observer would obtain high

agreement by chance and to insure that he really knows the system.

A further drawback is that neither of these methods accounts for chance agreement. It is for this reason that Scott developed a $\pi$ coefficient of agreement.

Scott's Coefficient. Scott (1955) argued that measures of simple percentage agreement may be inflated by chance agreement. He therefore proposed a $\pi$ coefficient which estimated the extent to which chance agreement has been exceeded when comparing two observers' scores:

$$\pi = \frac{P_o - P_e}{1 - P_e}$$

where $P_o = \frac{1}{n} \sum_{i=1}^{C} n_{ii}$

$n_{ii}$ = the number of items on which the two observers agreed

$n$ = the total number of items coded

and $P_e = \sum_{i=1}^{C} \left[ \sum_{j=1}^{J} n_{ij} / N.. \right]^2$ = proportion of agreement expected by chance

$n_{ij}$ = the number of codes in the $i$th category for the $j$th observer

$N..$ = the total number of codes across $C$ categories and $J$ observers

$P_e$ is based on the marginal distributions of categories across all observers, and the same $P_e$ is used for all comparisons. For each pairwise comparison of observers Scott assumed that their proportional distributions of marginals were symmetrical and approximately equal to the average proportional distributions of marginals obtained from all observers. Since it is not always feasible to meet this assumption, Cohen (1960) suggested an alternative procedure of calculating $P_e$.

Cohen's Kappa. Cohen (1960) proposed a $\kappa$ coefficient in which $P_e$, chance agreement, is based on observed marginal distributions rather than expected marginals and requires no assumption that marginals are symmetrical and proportional to known populational marginals. $P_o$, observer agreement, is computed in the same manner as Scott's, but $P_e$ is defined differently.

$$\kappa = \frac{P_o - P_e}{1 - P_e}$$

$$\text{and } P_e = \frac{1}{n^2} \sum_{i=1}^{C} n_{i+} \, n_{+i}$$

where $n_{i+}$ and $n_{+i}$ are observer marginals for each category and n is the number of items coded.

The reader should note that both Scott's $\pi$ and Cohen's $\kappa$ yield one agreement coefficient across two or more categories for each pair of observers. Application of these measures when using only one category is inappropriate.

Light's Extension of $\kappa$. Light (1971) agreed that $\kappa$ was conceptually attractive as a simple distance measure of agreement. He preferred, however, to view the $\kappa$ statistic as a distance measure of disagreement under the hypothesis of random agreement. That is,

$$\kappa = 1 - \frac{d_o}{d_e}$$

where $d_o = 1 - P_o$ = observed proportion of disagreement

and $d_e = 1 - P_e$ = expected proportion of disagreement.

Although Light's formulation is computationally equivalent to Cohen's $\kappa$, it conceptually facilitates extensions to other types of agreement measures such as agreement with more than two observers, comparisons of the joint agreement of several observers with a standard, measures of conditional agreement with two or more observers, and methods of

distinguishing patterns of agreement from levels of agreement between two observers. All of these methods require item by item comparisons, and most are computationally complex.

One of these measures, conditional agreement with two observers, allows one to compare agreement of each observer's score to a criterion score for "only those items which one observer /the expert/ placed in the ith specific category" (Light, 1971, p. 367). That is,

$$\kappa_{pi} = 1 - \left(\frac{1-n_{ii}}{n_{i+}} \Big/ 1 - \frac{n_{+i}}{n}\right)$$

where $n_{ii}$ = number of agreements between the observer and criterion coder on the ith category

$n$ = total number of items coded

$n_{+i}$ = marginal for the observer on the ith category

$n_{i+}$ = marginal for the criterion coder on the ith category

On advantage of $\kappa_p$, $\kappa$, and $\pi$ is that they can be tested for significance (see Fleiss, Cohen, & Everitt, 1969; Light, 1971). A limitation of these coefficients is that item by item comparisons are mandatory if they are to be used appropriately (Emmer, 1972). Since it is not always possible or convenient to obtain observational data in this form, Flanders (1967) has proposed a modification of Scott's $\pi$ using category totals.

Flanders' Modification. Flanders (1967) used the same formula as Scott, but modified the computation of $P_o$. Correction for chance agreement is computed separately for each pair of observers based on their observed marginals rather than using a common expected $P_e$ based on all observers as Scott did.

$$\pi_f = \frac{P_{of} - P_{ef}}{1 - P_{ef}}$$

where $P_{of} = 1 - \sum_{i=1}^{C} \left| \frac{f_{i1}}{f_{.1}} - \frac{f_{i2}}{f_{.2}} \right|$

$f_{i1}$ = the frequency for the ith category for observer 1

$f_{i2}$ = the frequency for the ith category for observer 2

$f_{.1}$ = the total frequency of codes across C categories for observer 1.

$f_{.2}$ = the total frequency of codes across C categories for observer 2

and where $P_{e_f} = \sum_{i=1}^{C} \left[ \left( \frac{f_{i1} + f_{i2}}{f_{.1} + f_{.2}} \right) / 2 \right]^2$.

The way in which Flanders has calculated $P_o$ is not affected by zero or low frequencies as is simple percentage agreement on category frequencies. However, his $P_o$ becomes dubious when two sets of ratings correlate highly. For example, if the first observer simply did not recognize half of the behavioral events occurring because he was a slow coder, while the second one did see them, it is possible that the following results could be obtained:

| | Observer 1 Frequency | Observer 2 Frequency |
|---|---|---|
| $C_1$ | 2 | 4 |
| $C_2$ | 6 | 12 |
| $C_3$ | 4 | 8 |
| $C_4$ | 5 | 10 |
| $C_5$ | 3 | 6 |

If Flanders' $\pi_f$ were utilized on these data, a coefficient of 1.00 would result· yet it is evident that the two observers do not perfectly agree on individual category comparisons. (See Appendix A)

While this type of agreement measure is appropriate if category proportions or percentages are employed in analysis, it would be a misleading measure if data were analyzed using category frequencies or three-stage patterns of behavior as a unit of analysis.

Garrett (1972) suggested an alternative procedure for calculating $\pi$ which is not biased by highly correlated data. $P_o$ is calculated in a manner similar to the aforementioned second method of simple percentage agreement. Her $P_e$ is computed in the same manner as Scott's $P_e$, but she considers the two observers to be the population each time.

One limitation with Garrett's $\pi_g$ is that it can cause an interpretation problem for infrequent occurrences of a given category that is similar to the one discussed for simple percentage agreement measures. It too, like $\pi_f$, can only be used as a descriptive statistic which cannot be subjected to a statistical test of significance.

Which Agreement Coefficient is Appropriate?

Although the conditions for measuring agreement have been specified, and a number of methods of obtaining agreement measures have been presented, the reader still may be left in a quandry--which observer agreement measure is most appropriate? Moreover, once an agreement coefficient has been selected, how large should that coefficient be in order to be acceptable? There are probably no two best answers to these questions. Based on the foregoing discussion, the previously specified conditions for measuring agreement, and the experience of the authors, the following procedures are recommended:

1. Observer agreement with a criterion. In the case of coding isolated unambiguous examples when making item by item comparisons, Cohen's $\kappa$ or Light's $\kappa_p$ seem most desirable. Scott's $\pi$ should generally be avoided, since

since it is often difficult to meet the assumption relative to marginal distributions. Cohen's $\kappa$ appears to be most appropriate for determining observer agreement on a scale (cluster of categories), providing that the scale is to be used as the unit of data analysis. On the other hand, if an individual category is the intended unit of analysis, Light's $\kappa_p$ seems most appropriate. If a sequential analysis of behavior patterns is planned (e.g., Collett & Semmel, 1971), $\kappa_p$ can be used for each pattern of interest. To be conservative, however, an agreement should be counted only if the order of recorded behaviors is identical for both observers for each instance of a particular pattern, disregarding any additional elements which are considered "noise" in the sequential vector.

Interpretation of these coefficients is yet another matter. As Light (1971) has noted, it is possible to test the significance of $\kappa$ and $\kappa_p$ if nominal data are used. Emmer (1972) has commented that such significance tests are seldom used with observer agreement measures--and perhaps for a good reason. To have demonstrated that an observer agrees with a criterion at a level significant beyond chance does not guarantee that such agreement is nearly perfect. It is assumed that, when Medley and Norton (1971) referred to nearly perfect agreement, they were expecting simple percentage agreement ($P_o$) of .85 to .90 or greater. If $P_o$ = .90 is considered to be a lower bound for making decisions on adequacy of observer skills, then $\kappa$, from the experience of the present authors, typically falls around .80 or greater, while $\kappa_p$ is less predictable.

Therefore, it seems most practical and logical to use the simple percentage agreement measure if $P_o \geq .85$ is demanded when making item by item comparisons of unambiguous events. Interpretation is further aided if an

observer is given a criterion test consisting of at least ten or more examples of each category (scale, chain) with an approximately equal number of examples of each unit of analysis.

If item by item comparisons are impossible or impractical, $\kappa$ and $\kappa_p$ can be modified by assuming that the lower of the total frequencies for each category for the two observers is the number of items upon which they agreed. While $\kappa$ and $\kappa_p$ (and $\pi$) can be used as a descriptive statistic in this manner, statistical tests of significance are clearly inappropriate (Emmer, 1972). Following the same line of reasoning, however, the aforementioned second method of simple percentage agreement appears most reasonable, if $P'_o \geq .85$ is demanded when comparing total category or scale frequencies of unambiguous events (assuming approximately equal representation of categories.)

2. Intraobserver agreement. It was also suggested that a measure of intraobserver agreement be taken on two observations of the same videotape segment of a realistic setting. It should be noted that the purpose of criterion-referenced agreement is to assure a trainer that his coders know the system nearly perfectly, while the intent of intraobserver agreement is to demonstrate that coders can apply their discrimination skills in an actual coding situation.

Since item by item comparisons are usually impractical for the latter measure, the modification of $\kappa$ discussed above and the two modifications of $\pi$ are possible choices for an overall measure of intraobserver agreement. However, since ambiguous events are likely to occur, acceptability of obtained coefficients should be interpreted more liberally than that for criterion-referenced agreement.

The present authors have used average simple percentage agreement with $\bar{P}'_o = .75$ as an acceptable lower limit of intracoder agreement when total

category frequencies are the unit of analysis and there are few low frequency categories. Alternatively, Flanders' method of calculating $P_o$ is not affected by low frequencies of categories as if $\overline{P}'_o$. His $P_o$ is preferable to $\overline{P}'_o$ when the distribution of category frequencies is unequal and some are quite low. The problem of a positive bias when pairs of ratings correlate highly is not as serious with the intraobserver check as it is with the criterion-referenced test, since ambiguous events are likely; and, more important, observers tend to see more and code more during the second viewing.

If Flanders' $\pi_f$ is used as an overall intraobserver measure, and if one sets $P_{of} = .75$ as minimally acceptable, then $\pi_f$ will typically fall around .65 to .70 for five- to fifteen-category systems.

## INVESTIGATOR ERRORS

While observer agreement is one of the first concerns of an investigator, reliabilities of observational records are more important. The plural of reliability is used because more than one intraclass correlation coefficient can often be calculated for a given set of observational records--particularly for multi-facet designs. Each coefficient estimates the extent to which elements of a certain facet in the design can be consistently discriminated from each other. Components of variance which are considered to be true sources of variation rather than error variance depend on the reliability coefficient of interest, which is in turn dependent on the design of the study and objectives of the investigation (Gleser, Cronbach, and Rajaratnam, 1965). Thus, the design of an observational study as well as the selection of facets which determine true and error variance components are important considerations, since these factors affect both the nature and degree of reliabilities of observational data.

### Accounting for Situational Factors

In response to these considerations, McGaw, et al., (1972) have contended that context or situational variables have been neglected or mistreated in design and analysis of observational studies. In disagreement with Medley and Mitzel (1963), they have purported that variance in teacher (and pupil) behavior from situation to situation may be more lawful than it is random. Although the situations were never clearly defined, it is assumed that context variables such as subject matter, class size, seating arrangements, group structure, nature of teacher and pupil task, time of the day (week), etc. were considered. McGaw and his associates have argued that situations should be treated as a separate facet in determining reliabilities. More specifically, assuming that an observable dimension of

teacher behavior is of interest, differences among teachers (T), and situations (S) are both considered to be true sources of variation. If situational factors are not taken into account, then variance attributable to situations is included in measurement error. It is the contention of McGaw, et al., that such designs or analyses which have treated situational variance as part of generic error variance may have limited the reliability with which teachers (or pupils) were discriminated.

Moreover, the interaction of teachers and situations (TxS) may be systematic and is likewise treated as a true source of variation. For example, it is sensible that in a large group social studies lesson with thirty pupils a teacher's questioning style may be quite different from his/her questioning style when interacting with one pupil working on an art project.

Components of variance which are treated as error include variance in behavior over occasions (O within TxS), variance attributable to rater differences (J), and other interaction components. That is, the total variance for their design,

$$\sigma_X^2 = \sigma_t^2 + \sigma_s^2 + \sigma_{ts}^2 + \sigma_\varepsilon^2$$

where $$\sigma_\varepsilon^2 = \sigma_{o(ts)}^2 + \sigma_j^2 + \sigma_{tj}^2 + \sigma_{sj}^2 + \sigma_{tsj}^2 + \sigma_{o(ts)j}^2 + \sigma_e^2$$

Thus, three coefficients of generalizability, or indices of reliability, may be estimated:

$$\beta_t^2 = \theta_t^2 \Big/ \left(\hat{\sigma}_t^2 + \hat{\sigma}_\varepsilon^2\right)$$

$$\hat{\rho}_s^2 = \hat{\sigma}_s^2 \Big/ \left(\hat{\sigma}_s^2 + \hat{\sigma}_\varepsilon^2\right)$$

$$\hat{\rho}_{ts}^2 = \hat{\sigma}_{ts}^2 \Big/ \left(\hat{\sigma}_{ts}^2 + \hat{\sigma}_\varepsilon^2\right)$$ [McGaw, et al., (1972), pp. 24-25].

The advantage of partitioning variance components and reliabilities in this manner is that data interpretation can be facilitated. For instance,

if both $\hat{\rho}_t^2$ and $\hat{\rho}_s^2$ turn out to be small, and $\hat{\rho}_{ts}^2$ is relatively large, it can be concluded that while neither differences in teachers (within situations) nor situations can be clearly discriminated, differences among teachers in their changes in behavior from one situation to another can be detected.

An obvious implication of such an approach is that situation effects should be considered when designing an observational study. To simplify data analysis procedures, it would appear advantageous that all teachers be observed an equal number of times in the same types of situations. Or, if such a design is unwieldy, all teachers might be observed an equal number of times in the same situation.

It should be noted that McGaw, et al., (1972) have treated the situations facet as a random factor in their design. If levels of situations are selected in a non-random manner, or are considered a fixed effect, generalizations must be restricted to only those levels rather than to the universe of situations as defined by McGaw, et al.

Finally, it is necessary that observation schedules include items or scales for recording situational elements if the situation is to be considered. Observation instruments used in Project PRIME (Kaufman, Semmel, & Agard, 1973) and by Medley and Norton (1971) are good examples of instruments which account for a number of situational variables.

Observer Assignment

It is well known that as the number of items on a test is increased, the reliability with which the test can discriminate subjects on a given dimension is likely to increase. The same principle applies to observational studies. For example, Medley and Mitzel (1958) found that increasing the number of observers per visit minimally affected reliability, while increasing

the number of visits per classroom substantially increased the reliability with which teachers were discriminated on five observation scales.

Medley and Mitzel (1963) suggested that a reliability check be performed in the field (before actual data collection) in order to estimate the number of visits necessary for an acceptable coefficient. More recently, however, Medley and Norton (1971) have concluded that such a procedure is of little value. Rather, as many visits per classroom as possible are desirable in collecting actual observational data.

Implications for observer assignment are that independent pairs of observers should ideally visit each classroom an equal number of times such that each rater is paired with every other rater equally often. Not only is analysis simplified by this procedure, but it is also possible to obtain a direct measure of observer agreement in the classroom by using intraclass correlation coefficients (Medley & Mitzel, 1958).

Sometimes it is impractical to assign pairs of observers to each classroom. In fact, in order to maximize the number of visits to each classroom, and to minimize observer effects, one observer per visit is most desirable providing that "each recorder observes each individual at least once, and observes every individual the same number of times. This number may vary across recorders--one recorder may see all individuals twice; another may see them three times each /Medley & Norton (1971); Exhibit 2, p. 17." With this procedure coefficients of observer agreement, stability of behaviors, and reliabilities of both individuals and groups can be calculated.

Poorly-designed Observation Systems

Another factor which can affect reliability is a poorly-designed observation system. From the authors experience, systems with high inference categories (i.e., poorly defined, or based on extremely subtle and/or complex cues) cause problems in observer agreement during training.

It logically follows that this can curtail reliability coefficients, and serious interpretation problems are likely to result.

In order to reduce such occurrences, Medley and Norton (1971) have stressed that "objectivity is ensured by defining categories so that these discriminations are based (1) on relatively obvious and easily recognized cues, and (2) on cues which are minimally dependent on sophisticated knowledge or on the observer's own set of values (p. 1)." Thiagarajan, M. Semmel, and D. Semmel (in press) have also delineated a method of concept analysis which can enhance objectivity of categories during development of a system or system-training materials.

Two recommendations have been proposed to those who use and/or are developing observation systems in order to help reduce such errors. First, Emmer (1972) has suggested that developers of observation systems, like developers of commercial tests, should include various reliability coefficients and note the nature of the sample studied when publishing their systems. Thus, other system users and developers would be able to make informed choices among observation systems and categories based on their reliability.

Secondly, Thiagarajan (1973) noted that a number of observation systems are highly dependent on their original developers for purposes of training coders. If persons other than the originators train observers on the systems, the intended nature of the systems can be inadvertantly distorted. This could be one reason why results from observational studies using the same system with different investigators have been inconsistent. (Soar, 1972). In order to reduce this problem, Thiagarajan has suggested that self-contained, criterion-referenced instructional packages be developed for observation systems. This procedure could enhance the consistency with which a system is used.

## SUMMARY

Observer disagreement is important insofar as it limits the reliabilities of observational records. Once a study is finished, the extent to which observer errors detract from reliabilities can be estimated by intraclass correlation coefficients providing certain design requirements are met (Medley & Mitzel, 1963). It is too late, however, at this point in time to retrain observers if observer disagreement is found to be seriously limiting reliabilities.

Hence, the previous discussion has evolved around methods and conditions under which observer agreement can be measured so as to minimize such an occurrence.

It has been concluded that observers should be trained to nearly perfect agreement with a criterion or expert coder on unambiguous examples of behavioral categories before actual data collection. Coders should then be expected to agree on unambiguous events encountered in the field. But disagreement on ambiguous events observed in the field should also be expected, since teachers and pupils do not always exhibit behaviors which neatly fall into predefined observational system categories. Disagreement on ambiguities may help reflect a more accurate representation of the real world (Medley and Norton, 1971).

Since the number of ambiguous events occurring in the field cannot be controlled, a measure of observer agreement in that situation is difficult to interpret. Rather, the best that can be done is to document that observers can accurately code unambiguous examples. This can be accomplished by showing observers a video tape containing only unambiguous examples.

In addition to criterion-related agreement, it was suggested that measures of intraobserver agreement be obtained by showing a video tape

twice to all observers in which conditions parallel those encountered in the field. The purpose of an intraobserver agreement measure is to demonstrate the extent to which each observer can consistently code under observational circumstances which closely approximate classroom conditions.

While several methods of calculating observer agreement have been proposed (e.g., Scott, 1955; Cohen, 1960; Flanders, 1967; Light, 1971; Garrett, 1972), little emphasis has been placed on interpretation of observer agreement measures. Due to lack of existing guidelines for making decisions on adequacy of observer training, relationships among various agreement measures were discussed. It was concluded that simple percentage agreement $\geq .85$ for each unit of data analysis was acceptable for a criterion-related measure of agreement on unambiguous videotaped examples. An overall proportion of agreement $\geq .75$ was also recommended for an intraobserver measure on a videotape representative of realistic classroom coding conditions.

In addition, the necessity of calculating agreement coefficients with the same type(s) of data (e.g., category frequencies, two-stage patterns, or scales) that are used in analysis of actual data collected in the study was emphasized.

While criterion-related and intraobserver agreement measures have been recommended for both before and during a study, these measures should not be used as evidence of observer agreement in the actual classroom. Rather these are measures to assist an investigator in documenting adequacy of observational skills. The purpose of such efforts are to minimize the possibility that observers are primarily responsible for potentially unreliable observational data.

After a study is finished reliabilities of observational data and coefficients of stability and observer agreement should be calculated by

using intraclass correlation coefficients. It was emphasized that there are many types of reliabilities, each depending on the design of the study and on how true and error variance components are partitioned.

Since teachers and pupils may behave differently in different situations, it was suggested that treatment of situations and subject X situation interactions as error variance in past studies may have limited the reliability with which teachers and pupils were discriminated. As a result, identification of classroom process variables that relate to pupil growth may have been obfuscated. It was recommended that situational factors be included in observation systems in an attempt to reduce the large amounts of error variance typically attributable to instability of human behavior. The identification of significant situational factors remains yet to be determined empirically, however.

It was also emphasized that appropriate methods of observer assignment can make it possible to determine the extent to which observer disagreement limits reliabilities relative to other sources of error such as instability of behavior across occasions. Finally, it was mentioned that poorly designed observation systems can also curtail reliabilities.

Considering the variety of different types of errors that can enter into observational studies, it is not surprising that few, if any, relationships among classroom process variables and pupil outcome measures have been yet established.

References


Cohen, J. A. A coefficient of agreement for nominal scales. Educational and psychological measurement, 1960, 20, 37-46.

Collett, L., & Semmel, M. I. The analysis of sequential classroom behavior. Unpublished paper, University of Michigan, Office of Research Services, School of Education, 1971.

Cronbach, L. J., Rajaratnam, M., & Gleser, G. Theory of generalizability: A liberalization of reliability theory. British Journal of Statistical Psychology, 1963, 16, 137-163.

Ebel, R. L. Estimation of the reliability of ratings. Psychometrika, 1951, 16, 407-424.

Emmer, E. Direct observation of classroom behavior. In N. Flanders and G. Nuthall, (eds.) The Classroom Behavior of Teachers, International Review of Education XVIII/1972/4 Special Number (pp. 508-528).

Flanders, N. A. Estimating reliability. In Amidon, E. J., and Hough, J. B. (eds) Interaction Analysis: Theory, research, and application. Reading, Mass.: Addison-Wesley, 1967, 161-166.

Fleiss, J. L., Cohen, J., & Everitt, B. S. Large sample standard errors of kappa and weighted kappa. Psych. Bulletin, 1969, 72, 323-327.

Garrett, C. S. Modification of the Scott coefficient as an observer agreement estimate for marginal form observation scale data. Center for Innovation in Teaching the Handicapped, Indiana University, Occasional Paper #6, 1972.

Haggard, E. A. Intraclass correlation and the analysis of variance. New York: Dryden Press, 1958.

Kaufman, M. J., Semmel, M. I., & Agard, J. A. Project PRIME: An Overview. United States Office of Education, Bureau of Education for the Handicapped, Division of Research, Intramural Research Program in conjunction with the Texas Education Agency, Department of Special Ed. and Special Schools, Division of Program Evaluation, 1973.

Light, R. J. Measures of response agreement for qualitative data: Some generalizations and alternatives. Psych. Bulletin, 1971, 76, (5), 365-377.

McGaw, B., Wardrop, J. L., & Bunda, M. A. Classroom observation schemes: Where are the errors? American Educational Research Journal, 1972, 9(11), 13-27.

Medley, D. M., & Mitzel, H. E. Application of analysis of variance to the estimation of the reliability of observations of teachers' classroom behavior. Journal of Experimental Education, 1958, 27, 23-35.

Medley, D. M., & Mitzel, H. E. Measuring classroom behavior by systematic observation. In N. L. Gage (ed.) Handbook of research on teaching. Chicago, Ill.: Rand-McNally, 1963, 247-328.

Medley, D. M., & Norton. The concept of reliability as it applies to behavior records. Prepared for the 1971 meeting of APA in Washington, D.C. (mimeo)

Scott, W. A. Reliability of content analysis: The case of nominal scale coding. Public Opinion Quarterly, 1955, 19, 321-325.

Semmel, M. I. Toward the development of a computer-assisted teacher training system (CATTS). In N. A. Flanders and G. Nuthall (eds.) The classroom behavior of teachers, International Review of Education XVIII/1972/4 Special Number (pp. 561-568).

Simon, A., & Boyer, E. G. Mirrors for Behavior II. Classroom Interaction Newsletter, Special Edition, Spring, 1970.

Smith, B. O. (ed.). Research in teacher education: A symposium. Englewood Cliffs, New Jersey: Prentice Hall, 1971.

Soar, R. Teacher behavior related to pupil growth. In N. Flanders and G. Nuthall, (eds.) The Classroom Behavior of Teachers, International Review of Education XVIII/1972/4 Special Number (pp. 508-528).

Thiagarajan, S. Instructional systems for interactional systems. Classroom Interaction Newsletter, 1973, 9(1), 13-22.

Thiagarajan, S., Semmel, M. I., & Semmel, D. S. Instructional development for training teachers of exceptional children: A sourcebook (in press).

# Appendix A

## Illustrations of Observer Agreement Measures

Two kinds of simple percentage agreement, Scott's $\pi$ (1955), Cohen's $\kappa$ (1960), Light's (1971) extension of $\kappa$, Flanders' (1967) modification of $\pi$, and two other measures of agreement between an observer and an expert will be considered respectively. In order to provide a thread of continuity among these methods, the same artificial data will be used in computational examples, although the data will appear in different "forms." Necessary assumptions, the required "form" of the data, the computational formula, examples, and advantages and limitations of each method will be given.

Assume that two observers coded a total of thirty events using a five category observation system. The categories used were:

$C_1$ = Teacher lecture
$C_2$ = Teacher comprehension question
$C_3$ = Teacher convergent question
$C_4$ = Teacher divergent question
$C_5$ = Teacher evaluative question

Following are some artificial data. The total frequencies of each category recorded during the thirty events are given for each observer:

| | Obs. 1 ($f_{i1}$) | Obs. 2 ($f_{i2}$) |
|---|---|---|
| $C_1$ | 4 | 2 |
| $C_2$ | 6 | 12 |
| $C_3$ | 4 | 8 |
| $C_4$ | 10 | 5 |
| $C_5$ | 6 | 3 |

Now, if observer agreement was measured on an item by item basis, there are two extreme cases, optimal and minimal agreement, in which the two observers' codes could result in the above total frequencies (marginals)

---

Insert Tables 1 & 2 Here

---

TABLE 1. An Optimal Case of Agreement on Individual Items Given Fixed Marginals

| Event | Observer 1 Code | Observer 2 Code | Agreement |
|---|---|---|---|
| 1 | $C_2$ | $C_2$ | A (=agree) |
| 2 | $C_2$ | $C_2$ | A |
| 3 | $C_1$ | $C_2$ | D (=disagree) |
| 4 | $C_4$ | $C_4$ | A |
| 5 | $C_4$ | $C_4$ | A |
| 6 | $C_5$ | $C_5$ | A |
| 7 | $C_5$ | $C_5$ | A |
| 8 | $C_5$ | $C_5$ | A |
| 9 | $C_5$ | $C_2$ | D |
| 10 | $C_4$ | $C_4$ | A |
| 11 | $C_1$ | $C_1$ | A |
| 12 | $C_5$ | $C_2$ | D |
| 13 | $C_1$ | $C_1$ | A |
| 14 | $C_4$ | $C_3$ | D |
| 15 | $C_2$ | $C_2$ | A |
| 16 | $C_3$ | $C_3$ | A |
| 17 | $C_3$ | $C_3$ | A |
| 18 | $C_3$ | $C_3$ | A |
| 19 | $C_4$ | $C_4$ | A |
| 20 | $C_2$ | $C_2$ | A |
| 21 | $C_2$ | $C_2$ | A |
| 22 | $C_5$ | $C_2$ | D |
| 23 | $C_1$ | $C_2$ | D |
| 24 | $C_4$ | $C_3$ | D |
| 25 | $C_4$ | $C_3$ | D |
| 26 | $C_4$ | $C_3$ | D |
| 27 | $C_4$ | $C_4$ | A |
| 28 | $C_2$ | $C_2$ | A |
| 29 | $C_4$ | $C_2$ | D |
| 30 | $C_3$ | $C_3$ | A |
| MARGINALS | Observer 1 | Observer 2 | $\Sigma A = 20$; $\Sigma D = 10$ |
| $C_1$ | 4 | 2 | |
| $C_2$ | 6 | 12 | |
| $C_3$ | 4 | 8 | |
| $C_4$ | 10 | 5 | |
| $C_5$ | 6 | 3 | |

TABLE 2. A Minimal Case of Agreement on Individual Items Given Fixed Marginals

| Event | Observer 1 Code | Observer 2 Code | Agreement |
|---|---|---|---|
| 1 | $C_2$ | $C_3$ | D (=disagree) |
| 2 | $C_2$ | $C_3$ | D |
| 3 | $C_1$ | $C_1$ | D |
| 4 | $C_4$ | $C_2$ | D |
| 5 | $C_4$ | $C_2$ | D |
| 6 | $C_5$ | $C_3$ | D |
| 7 | $C_5$ | $C_3$ | D |
| 8 | $C_5$ | $C_3$ | D |
| 9 | $C_5$ | $C_3$ | D |
| 10 | $C_4$ | $C_1$ | D |
| 11 | $C_1$ | $C_2$ | D |
| 12 | $C_5$ | $C_3$ | D |
| 13 | $C_1$ | $C_2$ | D |
| 14 | $C_4$ | $C_2$ | D |
| 15 | $C_2$ | $C_1$ | D |
| 16 | $C_3$ | $C_4$ | D |
| 17 | $C_3$ | $C_4$ | D |
| 18 | $C_3$ | $C_4$ | D |
| 19 | $C_4$ | $C_2$ | D |
| 20 | $C_2$ | $C_4$ | D |
| 21 | $C_2$ | $C_5$ | D |
| 22 | $C_5$ | $C_3$ | D |
| 23 | $C_1$ | $C_2$ | D |
| 24 | $C_4$ | $C_2$ | D |
| 25 | $C_4$ | $C_2$ | D |
| 26 | $C_4$ | $C_2$ | D |
| 27 | $C_4$ | $C_2$ | D |
| 28 | $C_2$ | $C_5$ | D |
| 29 | $C_4$ | $C_5$ | D |
| 30 | $C_3$ | $C_4$ | D |
| MARGINALS | Observer 1 | Observer 2 | $\Sigma A=0$; $\Sigma D=30$ |
| $C_1$ | 4 | 2 | |
| $C_2$ | 6 | 12 | |
| $C_3$ | 4 | 8 | |
| $C_4$ | 10 | 5 | |
| $C_5$ | 6 | 3 | |

Following are examples of different methods of calculating observer agreement using these data.

1. SIMPLE PERCENTAGE AGREEMENT: NOMINAL DATA

1.1. Assumption: Observational records must be analyzed on an item x item basis. On a given item or event observers either agree (A) or disagree (D).

1.2. Computational Formula

$$P_o = \frac{\Sigma A}{\Sigma A + \Sigma D}$$

where $\Sigma A$ = total number of agreeing pairs,

and $\Sigma D$ = total number of disagreeing pairs.

Range: $0 \leq P_o \leq 1.00$

1.3. Examples

Optimal Case: $P_o = \frac{20}{20 + 10} = \frac{20}{30} = .67$

Minimal Case: $P_o = \frac{0}{0 + 30} = .00$

1.4. Advantages

1.4.1. Computationally simple.

1.4.2. Discriminates between optimal and minimal cases.

1.5. Disadvantages

1.5.1. Does not account for "chance" agreement.

1.5.2. Can only be done for two observers at a time.

1.5.3. Often impractical to obtain nominal form observational data.

2. SIMPLE PERCENTAGE AGREEMENT: FREQUENCY DATA

2.1. Assumption: Observational data is analyzed on the basis of total tallies in each category. For a given category it is assumed that the lower score of two observers is the number of times they agreed, and the difference between scores is the number of times they disagreed.

2.2. Computational Formula

$P'_{o_i} = \frac{fi1}{fi2}$ or $\frac{fi2}{fi1}$, such that $0 \leq P'_{o_i} \leq 1.00$

where fi1 = score for observer 1 on the ith category

and fi2 = score for observer 2 on the same ith category

Also, $\overline{P}'_o = \frac{1}{C} \sum_{i=1}^{C} P'_{o_i}$ = average percent agreement

Range: $0 \leq \overline{P}'_o \leq 1.00$

2.3. Examples

Optimal Case: $P'_{o_2} = \frac{6}{12} = .50$

$$\overline{P}'_o = \frac{1}{5}\left(\frac{2}{4} + \frac{6}{12} + \frac{4}{8} + \frac{5}{10} + \frac{3}{6}\right) = .50$$

Minimal Case: Same as optimal case

2.4. Advantages

2.4.1. Computationally simple

2.4.2. Agreement for specific categories or a group of categories can be calculated.

2.5. Disadvantages

2.5.1. Does not differentiate between optimal and minimal cases

2.5.2. Can only be done for two observers at a time.

2.5.3. Does not account for "chance" agreement

2.5.4. May be affected by zero or low frequencies. For example:

| | Obs 1 | Obs 2 | Difference | $P'_{o_i}$ |
|---|---|---|---|---|
| $C_1$ | 1 | 2 | 1 | .50 |
| $C_2$ | 19 | 20 | 1 | .95 |

## 3. SCOTT'S π (1955)

Scott argued that measures of simple percentage agreement may be inflated by "chance" agreement. He therefore proposed a π coefficient which estimated the extent to which chance agreement has been exceeded.

### 3.1. Assumptions

3.1.1. Nominal, rather than frequency, data must be used. That is, on a given item two observers either agree or disagree.

3.1.2. Both observers must have identical marginal distributions of category proportions, equal to known populational values (Light, 1971, p. 367).

### 3.2. Using Contingency Tables

Light (1971) suggested that a contingency table is conceptually useful in understanding Scott's π, as well as Cohen's κ, and Light's extension of κ for measures of conditional agreement, agreement among more than two observers with each other and with a criterion, and measures of patterns of agreement. Therefore, a C x C contingency table will be constructed for purposes of illustrating the data.

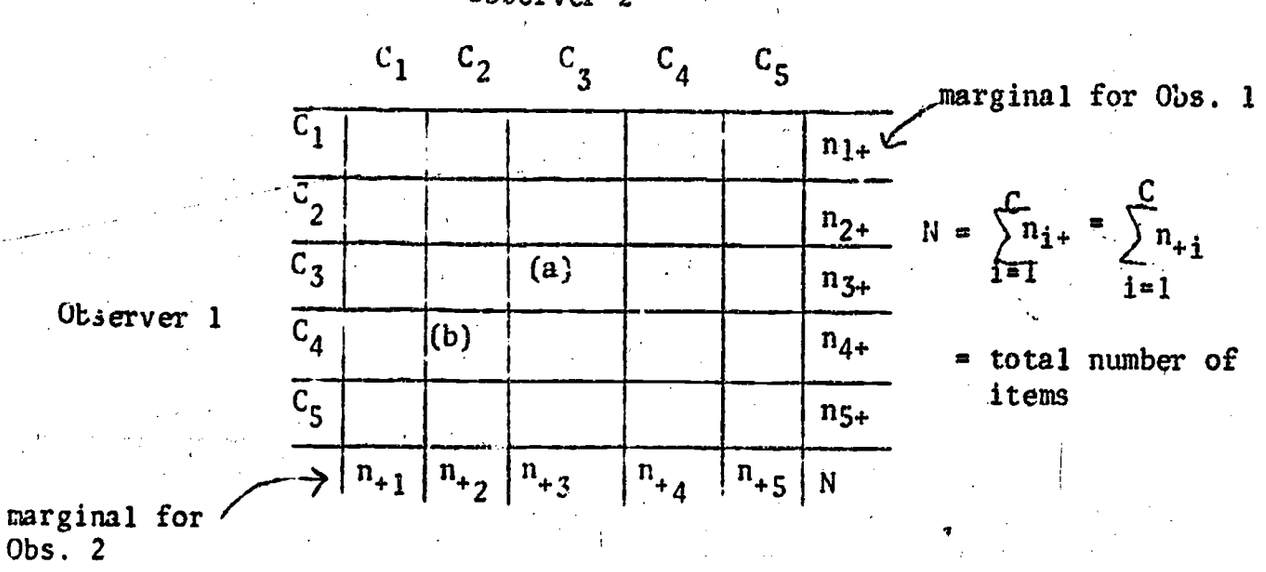


Figure 1. A 5x5 Contingency Table

For each behavioral event a tally can be made in the contingency table in Figure 1 showing agreement or disagreement and its nature. For example, suppose that for item (a) both observers recorded category #3. An entry would be tallied on the main diagonal ($C_{33}$), showing that the two observers agreed on this item. For item (b) the first observer recorded category #4, while the second observer saw it as category #2. An entry would be tallied off the main diagonal ($C_{42}$) showing how they disagreed.

For the optimal case the contingency table is:

| Observer 1 (Expert) \ Observer 2 | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $n_{i+}$ |
|---|---|---|---|---|---|---|
| $C_1$ | 2 | 2 | 0 | 0 | 0 | 4 |
| $C_2$ | 0 | 6 | 0 | 0 | 0 | 6 |
| $C_3$ | 0 | 0 | 4 | 0 | 0 | 4 |
| $C_4$ | 0 | 1 | 4 | 5 | 0 | 10 |
| $C_5$ | 0 | 3 | 0 | 0 | 3 | 6 |
| $n_{+i}$ | 2 | 12 | 8 | 5 | 3 | 30 |

For the minimal case the contingency table is:

| Observer 1 (Expert) \ Observer 2 | $C_1$ | $C_2$ | $C_3$ | $C_4$ | $C_5$ | $n_{i+}$ |
|---|---|---|---|---|---|---|
| $C_1$ | 0 | 4 | 0 | 0 | 0 | 4 |
| $C_2$ | 1 | 0 | 2 | 1 | 2 | 6 |
| $C_3$ | 0 | 0 | 0 | 4 | 0 | 4 |
| $C_4$ | 1 | 8 | 0 | 0 | 1 | 10 |
| $C_5$ | 0 | 0 | 6 | 0 | 0 | 6 |
| $n_{+i}$ | 2 | 12 | 8 | 5 | 3 | 30 |

## 3.3. Computational Formula

$$\pi = \frac{P_o - P_e}{1 - P_e}$$

where $P_o = \frac{1}{n}\sum_{i=1}^{C} n_{ii}$ = proportion of agreeing pairs

$n_{ii}$ = the number of items on which the two observers agreed for the ith category (main diagonal)

$n$ = total number of items coded

$$\text{and } P_e = \sum_{i=1}^{C} \left[ \sum_{j=1}^{J} n_{ij} / N.. \right]^2$$

$n_{ij}$ = the marginal for observer j for the ith category (there are J observers)

N.. = the total number of codes for all observers

3.4. Examples

Since the same $P_e$ is based on the population of observers and is used for each pairwise comparison of observers, it needs to be calculated only once. Suppose there are four observers in the study. Their marginals are:

| | Obs. 1 | Obs. 2 | Obs. 3 | Obs. 4 | $n_{i.}$ |
|---|---|---|---|---|---|
| $C_1$ | 4 | 2 | 6 | 7 | 19 |
| $C_2$ | 6 | 12 | 10 | 11 | 39 |
| $C_3$ | 4 | 8 | 2 | 1 | 15 |
| $C_4$ | 10 | 5 | 4 | 3 | 22 |
| $C_5$ | 6 | 3 | 8 | 8 | 25 |
| | | | | N.. = | 120 |

$$P_e = \left(\frac{19}{120}\right)^2 + \left(\frac{39}{120}\right)^2 + \left(\frac{15}{120}\right)^2 + \left(\frac{22}{120}\right)^2 + \left(\frac{25}{120}\right)^2 = .22$$

According to assumption 3.1.2., the proportional distributions of marginals for each pairwise comparison of observers should be symmetrical and they in turn are equal to known population proportions (population = J = 4 observers)

| | Population $[n_{i.}/N..]$ | | Observer 1 $[n_{i1}/n.1]$ | | Observer 2 $[n_{i2}/n.2]$ |
|---|---|---|---|---|---|
| $C_1$ | .16 | ↔ | .13 | ↔ | .07 |
| $C_2$ | .33 | ↔ | .20 | ↔ | .40 |
| $C_3$ | .12 | ↔ | .13 | ↔ | .27 |
| $C_4$ | .18 | ↔ | .33 | ↔ | .17 |
| $C_5$ | .21 | ↔ | .20 | ↔ | .10 |

For observers 1 and 2 it can be seen that their marginal distributions of category proportions are not very symmetrical nor are they equal to the population proportions. Thus, Scott's $\pi$ should not be used legitimately as a statistic with these data. For purposes of illustration $\pi$ will be computed in spite of the apparent violation of assumption 3.1.2.

Optimal Case

$$P_o = \frac{1}{30}(2 + 6 + 4 + 5 + 3) = \frac{20}{30} = .67$$

$$\pi = \frac{P_o - P_e}{1 - P_e} = \frac{.67 - .22}{1.00 - .22} = \frac{.45}{.78} = .58$$

Minimal Case

$$P_o = \frac{1}{30}(0 + 0 + 0 + 0 + 0) = 0.00$$

$$= \frac{0.0 - .22}{1.00 - .22} = -.28$$

3.5. Advantages

3.5.1. Corrects for chance agreement

3.5.2. Relatively simple to compute

3.5.3. Can be tested for significance if assumptions are met

3.5.4. Discriminates between optimal and minimal case

3.6. Disadvantages

3.6.1. Can only be used with two observers at a time

3.6.2. Assumes that observed marginals must be symmetrical and approximately equal to known population values, which may not always be the case.

4. COHEN'S $\kappa$ (1960)

4.1. Assumption: Nominal data are mandated. Identical marginals are not required, since chance agreement is based upon observed marginals.

4.2. Computational Formula

$$\kappa = \frac{P_o - P_c}{1 - P_e}$$

where $P_o = \frac{1}{N}\sum_{i=1}^{C} n_{ii}$ = proportion of agreeing pairs

and $P_e = \frac{1}{N^2}\sum_{i=1}^{C}(n_{i+})\ (n_{+i})$ = agreement expected by chance alone

$n_{ii}$ = number of items on which observers agreed for the ith category

$N$ = total number of items coded

$n_{i+}$ = marginal for observer 1 for the ith category

$n_{+i}$ = marginal for observer 2 for the ith category

4.3. Examples

Optimal Case (see contingency tables in Section 3.2.)

$$P_o = \frac{1}{30}\ (2+6+4+5+3) = \frac{20}{30} = .67$$

$$P_e = \frac{1}{(30)^2}\ (4x2 + 12x6 + 8x4 + 10x5 + 6x3) = \frac{180}{900} = .20$$

$$\kappa = \frac{.67 - .20}{1.00 - .20} = \frac{.47}{.80} = .59$$

Minimal Case

$$P_o = \frac{1}{30}\ (0+0+0+0+0) = .00$$

$$P_e = \frac{1}{(30)^2}\ (4x2 + 12x6 + 8x4 + 10x5 + 6x3) = \frac{180}{900} = .20$$

$$\kappa = \frac{0 - .20}{1.00 - .20} = \frac{-.20}{.80} = -.25$$

4.4. Advantages

4.4.1. Relatively simple to compute

4.4.2. Accounts for "chance" agreement

4.4.3. Possible to statistically interpret /see Cohen (1960) and Light (1971)/

4.4.4. Discriminates between optimal and minimal case

4.5. Disadvantages

4.5.1. Only good for two observers at a time

4.5.2. Cannot be legitimately used with quantitative data

4.5.3. Only measures overall agreement. Does not indicate agreement for specific categories, although surveyance of off-diagonal cells in the contingency table can be helpful.

## 5. LIGHT (1971): EXTENSIONS OF $\kappa$

Light agreed that $\kappa$ was conceptually attractive as a simple distance measure of agreement. He preferred, however, to view the $\kappa$ statistic as a distance measure of disagreement under the hypothesis of random agreement. That is:

$$\kappa = 1 - \frac{d_o}{d_e}$$

where $d_o = 1 - P_o$ = observed proportion of disagreement

and $d_e = 1 - P_e$ = expected proportion of disagreement

"Thus $\kappa$ becomes a ratio of measures of distance, or disagreements, between two observers, where distances are measured by counting up a series of ones and zeros. These distance measures are simply a function of the numbers of agreeing versus disagreeing pairs in the $2n$ total responses." (P. 367). It should be noted that Light's $\kappa$ is computationally equivalent to Cohen's $\kappa$.

Viewing $\kappa$ in these terms Light has extended it to:

1) conditional measures of agreement level with two observers,

2) measures of agreement among more than two observers,

3) measures of conditional agreement with more than two observers,

4) comparison of the joint agreement of several observers with a standard

and 5) a method of distinguishing patterns of agreement from levels of agreement between two observers.

Due to the computational complexity of most of these extensions of $\kappa$ they will not be discussed here. The interested reader is referred to Light (1971, pp. 367-376.)

## 6. LIGHT (1971): CONDITIONAL AGREEMENT, $\kappa_{p_i}$

### 6.1. Assumptions

The same assumptions hold as do those for Cohen's $\kappa$, except that $\kappa_{p_i}$ allows the comparison of each observer's score to a criterion (expert) score for only those items which the criterion placed in the $i$th specific category.

6.2. Computational Formula

$$\kappa_{p_i} = 1 - \left(\frac{1 - \frac{n_{ii}}{n_{i+}}}{1 - \frac{n_{+i}}{n}}\right)$$

where $n_{ii}$ = number of agreements between the observer and the criterion coder on the ith category

$n$ = total number of items coded

$n_{+i}$ = marginal for the observer on the ith category

$n_{i+}$ = marginal for the criterion coder on the ith category

6.3. Example

Optimal Case (for category 3; observer 1 is considered the expert)

$$\kappa_{p_3} = 1 - \frac{\left(1 - \frac{4}{4}\right)}{\left(1 - \frac{8}{30}\right)} = 1 - \frac{0.0}{.73} = 1.00$$

Minimal Case (for category 3)

$$\kappa_{p_3} = 1 - \frac{\left(1 - \frac{0}{4}\right)}{\left(1 - \frac{8}{30}\right)} = 1 - \frac{1.00}{.73} = -.37$$

6.4. Advantages

6.4.1. Can compare specific categories

6.4.2. Can test for significance

6.4.3. Accounts for chance agreement

6.4.4. Discriminates between optimal and minimal cases

6.5. Disadvantages

6.5.1. Cannot be used legitimately with quantitative data

7. FLANDERS' (1967) MODIFICATION OF SCOTT'S $\pi$

☞ Flanders used the general formula of $\pi$ (and $\kappa$) but modified the computation of $P_o$ for use with quantitative data. He also computed chance agreement similar to Scott's $P_e$ but used the average proportion per category rather than known populational proportions.

7.1. Computational Formula

$$\pi_f = \frac{P_{o_f} - P_{e_f}}{1 - P_{e_f}}$$

where $P_{o_f} = 1 - \sum_{i=1}^{C} \left| \frac{f_{i1}}{f_{.1}} - \frac{f_{i2}}{f_{.2}} \right|$

and $P_{e_f} = \sum_{i=1}^{C} \left[ \left( \frac{f_{ii}}{f_{.1}} + \frac{f_{i2}}{f_{.2}} \right) \frac{1}{2} \right]^2$

$f_{ii}$ = frequency for observer 1

$f_{i2}$ = frequency for observer 2

$f_{.1}$ = total number of tallies for C categories for observer 1

$f_{.2}$ = total number of tallies for C categories for observer 2

7.2. Example: (See Table 3)

7.3. Advantages

7.3.1. Uses frequency rather than nominal comparisons

7.3.2. Relatively simple to compute

7.3.3. $P_{o_f}$ is unaffected by low or zero frequencies in some categories as is average simple percentage agreement for frequency data

7.4. Disadvantages

7.4.1. Cannot be used legitimately for specific category agreement

7.4.2. Unable to statistically interpret

7.4.3. If two observational records correlate positively, Flanders' $\pi$ may overestimate observer agreement. This is most likely to happen when using an event-recording system rather than a unit-time recording system. For example, if one observer only detected half as many events as another, the following results could occur:

| | Obs. 1 | Obs. 2 |
|---|---|---|
| $C_1$ | 2 | 4 |
| $C_2$ | 6 | 12 |
| $C_3$ | 4 | 8 |
| $C_4$ | 5 | 10 |
| $C_5$ | 3 | 6 |
| $\Sigma f_i$ | 20 | 40 |

TABLE 3:

Flanders's Method of Calculating r for both Optional and Minimal Cases

| | Observer 1 (frequency, $f_{i1}$) | Observer 2 (frequency, $f_{i2}$) | (A) $\frac{f_{i1}}{f_{.1}}$ | (B) $\frac{f_{i2}}{f_{.2}}$ | \|(A) - (B)\| | Chance Agreement $\left[\frac{(A)+(B)}{2}\right]^2$ |
|---|---|---|---|---|---|---|
| 1 | 4 | 2 | $\frac{4}{30}$ = .13 | .07 | (.13-.07) = .06 | $\left(\frac{.13 + .07}{2}\right)^2$ = .01 |
| 2 | 6 | 12 | .20 | .40 | .20 | .09 |
| 3 | 4 | 8 | .13 | .27 | .14 | .04 |
| 4 | 10 | 5 | .33 | .17 | .16 | .06 |
| 5 | 6 | 3 | .20 | .10 | .10 | .02 |
| Sum | 30 | 30 | .99+ | 1.00 | (C) .66 | (D) .22 |

Flanders defines $P_{o_f}$ = 1 - (C) = 1 - .66 = .34

and $P_{e_f}$ = (D) = .22

$$r = \frac{P_{o_f} - P_{e_f}}{1 - P_{e_f}} \qquad r = \frac{.34 - .22}{1.00 - .22} = \frac{.12}{.78} = .15$$

Flanders' method would yield a coefficient of 1.00 with the above data. Yet it is apparent that the two observers do not "agree" on within category comparisons. This would not be necessarily undesirable if one were to analyze his data using proportions of frequencies in categories to the total number of recorded events.

However, if frequencies in each category were compared during data analysis, Flanders' method would yield a misleading overestimate of observer agreement.

## 8. A MODIFICATION OF SCOTT'S $\pi$ USING FREQUENCY DATA

In order to overcome the bias of Flanders' method when data are positively correlated and when using event recording systems, Garrett (1972) suggested a modification of Scott's $\pi$ that is unaffected by correlated observational records.

### 8.1. Computational Formula

$$\pi_g = \frac{P_{o_g} - P_{e_g}}{1 - P_{e_g}}$$

where $$P_{o_g} = \frac{1}{C} \sum_{i=1}^{C} \frac{\min}{\max} \{f_{i1}, f_{i2}\} / \{f_{i1}, f_{i2}\}$$

and $$P_{e_g} = \sum_{i=1}^{C} \left| \frac{f_{i1} + f_{i2}}{f_{.1} + f_{.2}} \right|^2$$

$P_{o_g}$ is equivalent to $\bar{P}_o'$ in Section 2. $P_{e_g}$ is equivalent to Scott's $P_e$ in Section 3.

### 8.2. Example: (See Table 4)

### 8.3. Advantages

8.3.1. Computationally simple

8.3.2. Unaffected by correlated ratings

8.3.3. Uses frequency rather than nominal data, which is helpful when item x item comparisons are impractical or impossible to obtain.

8.3.4. Corrects for chance agreement, based on observed marginals.

TABLE 4 : Garrett's Modification of Scott's $\pi$ for Both Optimal and Minimal Cases

| | Observer 1 $f_1$ | Observer 2 $f_2$ | $\frac{\min f_{i1}, f_{i2}}{\max f_{i1}, f_{i2}}$ | $\left(\frac{f_{i1} + f_{i2}}{f_{.1} + f_{.2}}\right)^2$ |
|---|---|---|---|---|
| 1 | 4 | 2 | $\frac{2}{4} = .50$ | $\left(\frac{6}{60}\right)^2 = .01$ |
| 2 | 6 | 12 | .50 | .09 |
| 3 | 4 | 8 | .50 | .04 |
| 4 | 10 | 5 | .50 | .06 |
| 5 | 6 | 3 | .50 | .02 |
| Sum | 30 | 30 | 2.50 | .22 |

$$P_{o_g} = \frac{1}{5}(2.50) = .50$$

$$P_{e_g} = .22$$

$$\pi_g = \frac{.50-.22}{1.00-.22} = \frac{.28}{.78} = .36$$

8.4. Disadvantages

8.4.1. Measures overall agreement rather than specific category agreement

8.4.2. Affected by low or zero frequencies of categories

8.4.3. Cannot statistically interpret

TABLE 5. SUMMARY CHART OF OBSERVER AGREEMENT MEASURES USING THE SAME ARTIFICIAL DATA FOR C CATEGORIES (C≥2)

| METHOD | TYPE OF DATA REQUIRED | FORMULA | OPTIMAL CASE | MINIMAL CASE | RANGE |
|---|---|---|---|---|---|
| Simple Percentage | Nominal | $P_o = \frac{\Sigma A}{\Sigma A + \Sigma D}$ | .67 | .00 | $0.0 \leq P_o \leq 1.00$ |
| Simple Percentage | Frequency | $\overline{P}'_o = \frac{1}{C}\sum_{i=1}^{C} \frac{\min\{fi1, fi2\}}{\max\{fi1, fi2\}}$ | .50 | .50 | $0.0 \leq \overline{P}'_o \leq 1.00$ |
| Scott | Nominal | $\pi = \frac{P_o - P_e}{1 - P_e}$ | .58 | -.28 | $\approx -.50 \leq \pi \leq 1.00$ |
| Cohen | Nominal | $\kappa = \frac{P_o - P_e}{1 - P_e}$ | .59 | -.25 | $\approx -.50 \leq \kappa \leq 1.00$ |
| Light | Nominal | $\kappa = 1 - \frac{d_o}{d_e}$ | .59 | -.25 | $\approx -.50 \leq \kappa \leq 1.00$ |
| Flanders | Frequency | $\pi_f = \frac{P_{o_f} - P_{e_f}}{1 - P_{e_f}}$ | .15 | .15 | $\approx -1.00 \leq \pi_f \leq 1.00$ |
| Garrett | Frequency | $\pi_g = \frac{P_{o_g} - P_{e_g}}{1 - P_{e_g}}$ | .36 | .36 | $\approx -.50 \leq \pi_g \leq 1.00$ |

TABLE 6. Summary Chart of Observer Agreement Measures Using the Same Artificial Data for One Category ($C_3$)

| METHOD | TYPE OF DATA REQUIRED | FORMULA | OPTIMAL CASE | MINIMAL CASE | RANGE |
|---|---|---|---|---|---|
| Simple Percentage | Frequency | $P'_{Oi} = \frac{\min\{fi1, fi2\}}{\max\{fi1, fi2\}}$ | .50 | .50 | $0.0 \leq P'_{Oi} \leq 1.00$ |
| Light | Nominal | $\kappa_{Pi} = 1 - \frac{\left(1-\frac{nii}{ni+}\right)}{\left(1-\frac{n+i}{n}\right)}$ | 1.00 | -.37 | $\kappa_{Pi} \leq 1.00$ |